XML for Technical Reports

Kurt Maly, M. Zubair (maly,zubair)@cs.odu.edu)

Old Dominion University

Norfolk, VA, 23529

http://dlib.cs.odu.edu

Outline

- NISO Z39.18 and prototype DTD
- Future Directions for Z39.18
- Other XML related projects at ODU

ANSI/NISO Z39.18-1995

Scientific and Technical Reports – Elements, Organization, and Design

Z39.18 Scope

- Teaches best practices
  - Structure
  - Content
  - Uniformity
  - Bibliographic information
- Teaches format
  - Style
  - Relations
- Teaches presentation methods
  - Visual and tabular matter
  - Equations
  - Paginating
  - Printing

Audience

- More geared towards authors of reports than readers (i.e., resource discovery)
- Librarian happy because of good bibliographic information in well known parts of report (Title page)
- More geared towards paper and ink reports than electronic dissemination and presentation

Demo

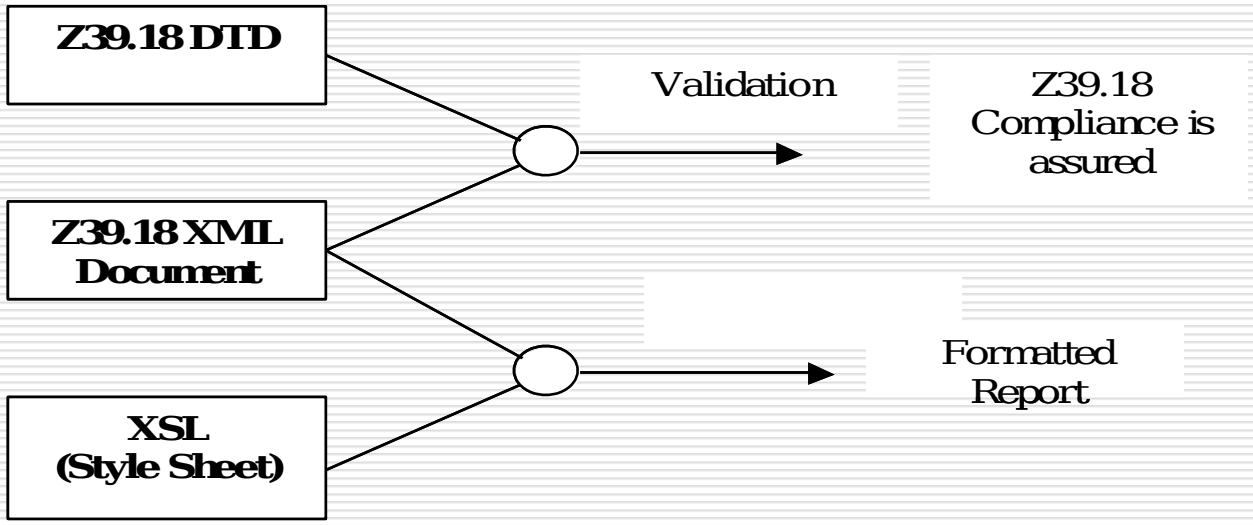
Demonstrate report: z38.19.pdf

What has XML to do with its revision?

- New standard geared also towards electronic dissemination, preservation, and discovery
- Clear separation of data and metadata
- Intended transport http(web)

Demonstration

Presentation in digital format



Demonstration

Document Type Definition (DTD)

The DTD provides definition of the structure of the Z39.18 XML document and the hierarchy of elements, their order of appearance, and constraints of how many times they should appear.

Show sample: z39.18.dtd

Demonstration

XSL (Style Sheet)

The XSL) as used within the Z39.18 context provides a mechanism for presentation of the data available in the XML document.

It provides formatting information, ordering of the presentation (need not be the same order as in the XML document) and can generate extra metadata such as table of contents, list of figures,

Multiple XSL sheets can be used for the same document to accommodate the needs of various communities. For example a style sheet can be provided for web publishing of reports, another for printed reports.

Show sample: z39.18.xsl

Demonstration

XML Document

The XML document contains the Z39.18 report along with its metadata. The elements in the XML document should comply to the DTD provided and which will be used to validate the XML document.

Show sample: sample.xml

Demonstration

Show report of sample.xml with z39.18.xsl applied

Show sample: sample.html

Commercial Tools

- Plug-Ins to existing word processors (Microsoft Word)
- Stand Alone XML Editors

Extyles - Inera

- Helps in creating XML document based on a specified DTD in the familiar Microsoft Word interface
- Support for the complete publication workflow process (Editing, Proof and Typset Corrections, Print and Create PDFs, etc.)

URL: http://www.inera.com

i4I – x4o

x4o allows you to

- create XML content based on a specified DTD in the familiar Microsoft Word interface
- create custom DTDs and XML templates based on specified DTDs.

URL: http://www.i4i.com/x4o.htm

Standalone Tools

Few Examples:

- ADEPT http://www.arbortext.com/
- XML Spy http://www.xmlspy.com/
- Amaya http://www.w3.org/Amaya/
- Xeena http://www.alphaworks.ibm.com/tech/xeena

Future Directions

- Address pending issues and take initial Z39.18 DTD to the next level. Collaborate with existing efforts like Docbook.

- Batch Processing for existing corpus (Converting into XML documents) and building of high level services.

DTD Issues – Handling Equations

Few models in use by several publishers:12083, Elsevier,

MathML, and TeX. (Nature: ISO12083,1994; Blackwell: MathML; IEEE: Tex)

MathML, unlike 12083 math, which is strictly presentation markup, can be used for presentation or content markup (expose underlying mathematical structure of an expression).

Neither 12083 math nor MathML can be natively displayed in most current browsers.

Current Solution: Convert equations into image usually in GIF format (Archon Project: http://archon.cs.odu.edu).

Handling of Chemical Formulas

DTD Issues – Handling Tables

CALS model: In use by several publishers DTD, though modified differently.

The CALS model is based on the MIL-M-38784B 910201 DTD originally developed for the US Department of Defense.

Docbook also uses CALS model.

DocBook is general purpose [XML] and [SGML] document type particularly well suited to books and papers about computer hardware and software (though it is by no means limited to these applications).

DTD Issues – Linking

- Inra-Document Links (figure citations, equation citation, table citation, reference citation to reference in bibliography, footnote citation, etc.

- Outside Links
  - Bibliographic links (CERN, ODU Archon: Demo, Open URL)

  - External Database: Accessed by standard format numbers for which links can be created. For example Genbank (http://www.ncbi.nlm.nih.gov/) is the NIH genetic sequence database and it holds an annotated collection of all publicly available DNA sequences.
  - Supplementary Material

DTD Issues – Collaboration

Work with existing effort like Docbook:

http://www.oasis-open.org/docbook/specs/cs-docbook-docbook-4.2.html

Docbook addresses a number of common issues.

Converting Existing Corpus

Need for batch processing tools with some human intervention that can convert existing corpus into structured XML documents that are consistent with Z39.18 DTD.

These documents then can be searched and processed electronically.

The process should be cost-effective with high accuracy

ODU is working in developing PDF extraction tools that can lead to creation of XML documents from scanned documents in PDF format.

High Level Services

Once we have publications in electronic format, a number of high level services can be supported, for example:

Annotation and review support

Cross citation and reference linking

Equation based search

Demo: Archon project features.

Sample of Digital Library Projects at ODU

- Archon:This project is building an Open Archives Initiative compliant federated digital library with an emphasis on physics for the National Science, Mathematics, Engineering, and Technology Education Digital Library (Sponsor: NSF).
- Kepler: framework that gives publication control to individual publishers, support speedy dissemination, and addresses interoperability. (Sponsor: NSF)
- Technical Report Interchange: Collaborative effort between NASA Langley Research Center, Los Alamos National Laboratory, Air Force Research Laboratory, Sandia National Laboratory and Old Dominion University to enable integration of technical reports. (Sponsor: NASA, LANL, SANDIA)

XML is the key technology used for these projects